HMFlow: Hybrid Matching Optical Flow Network for Small and Fast-Moving Objects

Suihanjin Yu, Youmin Zhang, Chen Wang, Xiao Bai, Liang Zhang
School of Computer Science and Engineering
Beihang University
Beijing, China
Email: {fakecoderemail, youmi, wangchenbuaa, baixiao, liang.z}@buaa.edu.cn

Edwin R. Hancock
Department of Computer Science
University of York
York, UK
Email: erh@cs.york.ac.uk

***Abstract*—In optical flow estimation task, coarse-to-fine (C2F) warping strategy is widely used to deal with the large displacement problem and provides efficiency and speed. However, limited by the small search range between the first images and warped second images, current coarse-to-fine optical flow networks fail to capture small and fast-moving objects which disappear at coarse resolution levels. To address this problem, we introduce a lightweight but effective Global Matching Component (GMC) to grab global matching features. We propose a new Hybrid Matching Optical Flow Network (HMFlow) by integrating GMC into existing coarse-to-fine networks seamlessly. Besides keeping in high accuracy and small model size, our proposed HMFlow can apply global matching features to guide the network to discover the small and fast-moving objects mismatched by local matching features. We also build a new dataset, named Small and Fast-Moving Chairs (SFChairs), for evaluation. The experimental results show that our proposed network achieves considerable performance, especially at regions with small and fast-moving objects.**

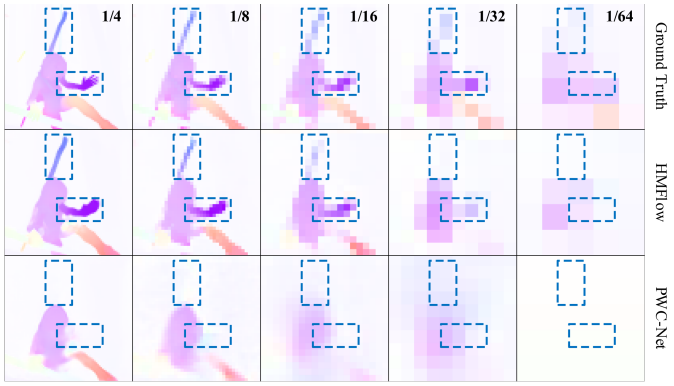

Fig. 1. The Flow Spatial Pyramid. The swinging arm and slender stick in blue boxes are small and fast-moving objects.

I. Introduction

Optical flow estimation plays an important role in many computer vision tasks, such as video object segmentation [1], [2], action recognition [3], [4], and autonomous driving [5], [6].

Traditional methods typically estimate optical flow by energy minimization in a coarse-to-fine (C2F) framework [7], [8], [9], [10]. The early deep learning based end-to-end optical flow networks [11], [12] are based on encoder-decoder architecture, which possess strong flexibility with large size of model parameters, causing high computing cost. Recently, the coarse-to-fine architecture networks [13], [14], [15] readopt traditional coarse-to-fine warping strategy [7] to provide accuracy and speed. These networks achieve high performances with relatively small model sizes, while they also inherit the problem of capturing the small and fast-moving objects from traditional coarse-to-fine methods.

The main problem lies in the contradiction between limited search range of warping based local matching and large displacement of small objects. To deal with large displacement problem efficiently, the coarse-to-fine warping strategy only sets a very limited search range at all resolution levels and matches between the first images and warped second according to up-sampled flows estimated at previous resolution levels. For a small and fast-moving object whose relative motion is larger than its own scale [16], if it disappears at the low-resolution levels, its location in the warped high-resolution image will shift with the background. Once it offsets out of the local search range, the mismatching occurs. And the flow estimation will be wrong at these regions at the subsequent high-resolution levels.

To solve the aforementioned problems, in this paper, we propose an approach to enhance the performance of coarse-to-fine networks at small and fast-moving object regions. Meanwhile, the proposed network only introduces a few network parameters but keep both high accuracy and computing efficiency. The key idea is to guide the coarse-to-fine network by global matching information for long-distance matching. A lightweight Global Matching Component (GMC) is proposed to calculate global matching features without image warping. To achieve our goal, we build Hybrid Matching Optical Flow Network (HMFlow) by integrating GMC into coarse-to-fine network. In conjunction with the limited local matching features of coarse-to-fine network, the global matching features of GMC can reduce the reliance of up-sampled flows, and guide the network discover the flows of small and fast-moving objects at the resolution levels which can distinguish them. And the high-quality warping based local matching features can make the HMFlow keep high accuracy. Fig. 1 demonstrates the performance of HMFlow compared to PWC-

Net [15]. The flows of swinging arm and slender stick, which are missed by PWC-Net, can be recovered by HMFlow as they appear at $1/16$ resolution level.

We build a new dataset for small and fast-moving objects, named Small and Fast-Moving Chairs (SFChairs). The scenes in SFChairs include small and fast-moving foreground objects and slow-moving background. The profiles and motions of every foreground objects are recorded for quantitative and qualitative evaluations. The effectiveness of network is borne out by experiments on this dataset.

We summarize our contributions as follows. First, we propose a lightweight but effective Global Matching Component (GMC) to produce global matching features that can cover the motion range of small and fast-moving objects. Second, we propose a Hybrid Matching Optical Flow Network (HMFlow), which integrates GMC to capture the small and fast-moving objects. Our proposed network still keeps a lightweight model size and high accuracy. Third, we build a specific dataset, SFChairs, for flow estimation evaluation, especially for small and fast-moving object regions. Experiments based on this dataset show both quantitative and qualitative results.

II. Related Work

Traditional Methods. The work of Horn and Schunck [17] is the pioneer for optical flow estimation. This method estimates the optical flow by optimizing an energy function based on brightness constancy and spatial smoothness assumptions. Several works [18], [19], [20] follow this pipeline.

To efficiently solve the large displacement problem, several works [7], [8], [9], [10] using coarse-to-fine (C2F) warping strategy are introduced. This strategy estimates flows in a multi-resolution spatial pyramid. An initial coarse flow is estimated and rectified in higher resolution by local matching between the first image and the warped second image according to coarser flow. However this scale-based methods is doom to fail in small and fast-moving objects due to their reliance on low-resolution estimation results. Xu *et al*. [21] use the SIFT descriptors to match these objects and extend the initial flow accordingly. Thomas *et al*. [22], [16] introduce sparse discriptor matching to guide the variational method to address this problem. Hu *et al*. [23] estimates the flow of small and fast-moving objects by setting and matching new pixel seeds at different levels. All of the above solutions rely on the introduction of global matching information. But they are not differentiable and cannot be easily trained in an end-to-end manner. Our GMC is inspired by them and realized by neural convolutional network.

Deep Learning Methods. FlowNet [11] and FlowNet 2.0 [12] are encoder-decoder architecture based end-to-end optical flow networks. The range of corresponding features they search for is the size of receptive field. By first encoding and then decoding, they expand receptive field to whole images and achieve global matching without warping according to low-resolution flows. The encoder-decoder architecture performs better than coarser-to-fine at dealing with flow estimation of small and fast-moving objects [11]. However, it also has to train a model with very large size.

Based on coarse-to-fine architecture, recent end-to-end optical flow networks [13], [14], [15] can get the accuracy on par with encoder-decoder networks but extremely reduce the model size. SPyNet warps the second image according to coarser flows, and then refined the estimated flows by convoluting the first and the warped image. Rather than warping on image, LiteFlowNet, PWC-Net and their varieties [24], [25], within a pretty small search range (e.g., only 4 neighboring pixels [15], [14]) at each resolution level, calculate their matching cost with cost volume upon the first and warped image features. However, all these models fail on small and fast-moving objects because of their limited searching range within coarse-to-fine architecture [15], [26]. Compared with above-mentioned methods, our HMFlow with lightweight GMC can achieve high accuracy while keeping a small size of the model.

Devon [26] avoids using spatial pyramid and warping, and gives much more accurate estimation of the small and fast-moving objects than [14], [15]. But it maintains full resolution at all stages and causes huge memory consumption. Our HMFlow with complete coarse-to-fine architecture is more memory friendly and can achieve higher accuracy.

Datasets. The common public datasets for the evaluation of optical flow methods, such as FlyingChairs [11], FlyingThings3D [27], MPI Sintel [28], KITTI 2012 [29], and KITTI 2015 [5], just concentrate on general problems of optical flow estimation. Some specialized datasets are built for special problems. Flying Vehicles with Rain (FVR) [30] is an optical flow dataset for rainy scenes. RoamingImages [31] is for multi-frame optical flow with occlusions. However, there is no specific dataset for evaluating the performance on regions with small and fast-moving objects. In this paper, we build a new synthetic dataset, by which the quantitative and qualitative evaluation for this problem can be done.

III. Approach

Our approach solves the problem of capturing the small and fast-moving objects of end-to-end coarse-to-fine (C2F) networks. We introduce global matching features to mitigate the short of limited search range of C2F network's warping based local matching. We design a lightweight Global Matching Component (GMC) to learn global matching features. We combine the global and local matching features of GMC and C2F network into hybrid matching features, and build Hybrid Matching Optical Flow Network (HMFlow). The hybrid matching features can guide network to capture these missed objects and keep network's high accuracy.

A. Matching And Estimating

The key of optical flow estimation is matching. From this perspective, the end-to-end optical flow networks can be divided into two parts called the feature matching part and the flow estimating part. The former computes matching features, and the latter estimates optical flow with matching features.

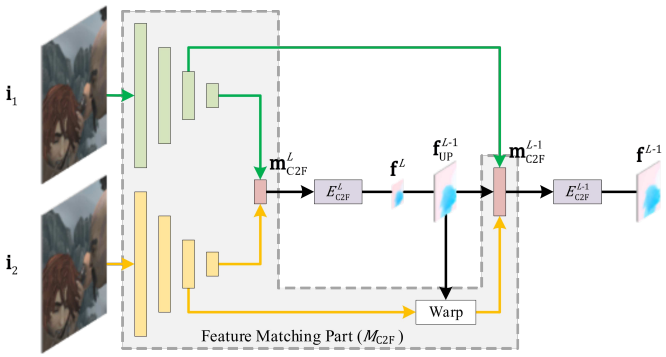

Fig. 2. The Function Partition of C2F Network. The feature matching part M_{C2F} is in the grey region, the flow estimating part E_{C2F} is the purple blocks.

The feature matching part M is used to calculate the matching features. We define this part as follow:

$$\mathbf{m}^0, \cdots, \mathbf{m}^l, \cdots, \mathbf{m}^L, \mathbf{u}^0, \cdots, \mathbf{u}^l, \cdots, \mathbf{u}^L = M(\mathbf{i}_1, \mathbf{i}_2) \quad (1)$$

where $\mathbf{i}_1$ and $\mathbf{i}_2$ are two consecutive images, L is the max spatial pyramid level, the superscript $\cdot^l$ means the feature's resolution is $1/2^l$ of $\mathbf{i}_1$ and $\mathbf{i}_2$, $\mathbf{m}^l$ and $\mathbf{u}^l$ are the output matching feature, which associates pixels in $\mathbf{i}_1$ with their corresponding pixels in $\mathbf{i}_2$, and unmatching feature at $1/2^l$ resolution level.

The flow estimating part E estimates flows according to the outputs of M. We define this part at $1/2^l$ resolution level as follows:

$$\mathbf{f}^l = E^l(\mathbf{m}^l, \mathbf{u}^l) \quad (2)$$

where $\mathbf{f}^l$ is estimated flow at $1/2^l$ resolution level.

The function partition of C2F network is shown in Fig. 2. The feature matching part M_{C2F} in the grey region includes a Siamese network filled by green and yellow, and red matching layers. The matching layers, like warp [13], [14] and cost volume [14], [15] produce matching features $\mathbf{m}^l_{\text{C2F}}$. The flow estimating part E^l_{C2F} in the purple block is the convolution layers between the matching feature and estimated optical flow at $1/2^l$ resolution level. By convolutional neural network, E^l_{C2F} can directly obtain the refined flow $\mathbf{f}^l$ according to the warping based local matching feature $\mathbf{m}^l_{\text{C2F}}$ and the up-sample flow $\mathbf{f}^l_{\text{UP}}$ without the need to estimate the residual flow first [15], [25].

For C2F networks, the matching features $\mathbf{m}^l_{\text{C2F}}$ are the only matching information for the corresponding flow estimating part E^l_{C2F}. For an L-level pyramid setting, C2F networks only need a partial matching cost with a limited search radius of r pixels. A one-pixel motion at the top level corresponds to 2^{L-1} pixels at the full resolution images. Thus the r are set to be small.

Fig. 3 demonstrates that the range that $\mathbf{m}^l_{\text{C2F}}$ can cover decreases as resolution increases. Hence matching process must depend on the warped features according to up-sampled flows to deal with large displacement. This design comes from traditional methods and leads to the same problem of capturing the small and fast-moving objects.

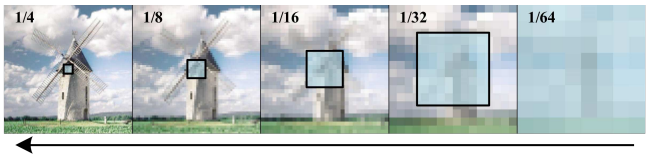

Fig. 3. The Local Search Range of C2F Network. In this sample, the resolution of inputs is 512×512, the search radius is $r = 4$ pixels [15], the blue region is the search range relative to inputs' resolution at each resolution level. The matching order of network is in arrow direction.

TABLE I
GLOBAL MATCHING COMPONENT (GMC)

	Layer	Kernel	Str.	Ch.	Input	Output
Encoding	Conv1A	7×7	2	c_1	$\mathbf{i}_1 \| \mathbf{i}_2$	$\mathbf{c}^1_{\text{A}}$
	Conv1B	7×7	1	c_1	$\mathbf{c}^1_{\text{A}}$	$\mathbf{c}^1_{\text{B}}$
	Conv2A	5×5	2	c_2	$\mathbf{c}^1_{\text{B}}$	$\mathbf{c}^2_{\text{A}}$
	Conv2B	5×5	1	c_2	$\mathbf{c}^2_{\text{A}}$	$\mathbf{c}^2_{\text{B}}$
	Conv3A	3×3	2	c_3	$\mathbf{c}^2_{\text{B}}$	$\mathbf{c}^3_{\text{A}}$
	Conv3B	3×3	1	c_3	$\mathbf{c}^3_{\text{A}}$	$\mathbf{c}^3_{\text{B}}$
	...	...	...	...	...	...
	Conv$l$A	3×3	2	c_l	$\mathbf{c}^{l-1}_{\text{B}}$	$\mathbf{c}^l_{\text{A}}$
	Conv$l$B	3×3	1	c_l	$\mathbf{c}^l_{\text{A}}$	$\mathbf{c}^l_{\text{B}}$
	...	...	...	...	...	...
	Conv$L$A	3×3	2	c_L	$\mathbf{c}^{L-1}_{\text{B}}$	$\mathbf{c}^L$
Decoding	Match$L$A	3×3	1	c_L	$\mathbf{c}^L$	$\mathbf{m}^L_{\text{G}}$
	Deconv$L$-1	4×4	2	c_{L-1}	$\mathbf{m}^L_{\text{G}}$	$\mathbf{d}^{L-1}$
	Match$L$-1	3×3	1	c_{L-1}	$\mathbf{d}^{L-1} \| \mathbf{c}^{L-1}_{\text{B}} \| \mathbf{f}^{L-1}_{\text{UP}}$	$\mathbf{m}^{L-1}_{\text{G}}$
	...	...	...	...	...	...
	Deconv$l$	4×4	2	c_l	$\mathbf{m}^{l-1}_{\text{G}}$	$\mathbf{d}^l$
	Match$l$	3×3	1	c_l	$\mathbf{d}^l \| \mathbf{c}^l_{\text{B}} \| \mathbf{f}^l_{\text{UP}}$	$\mathbf{m}^l_{\text{G}}$

[a] The **Str.** and **Ch.** indicate the Stride and Output Channels of convolution layers.

B. Global Matching Component (GMC)

To capture the small and fast-moving objects, traditional C2F methods introduce new global sparse descriptor matching to discover the missed objects. But they are not differentiable and cannot be easily trained in an end-to-end manner. The image-guide filter networks [32], [33] use image features to guide networks to discover more details. But its not enough to recover the motion of the objects moving alone, because these image features can just provide with structure information, rather than the matching information. The encoder-decoder optical flow networks [11], [12] can produce global matching features by expanding receptive field, but their model sizes are too large to be used for small optical flow networks.

Inspired by the above work, we design a Global Matching Component (GMC). The GMC is a lightweight U-Net [34] encoder-decoder network, which belongs to feature matching part M_{GMC} functionally. It focuses on providing E^l_{C2F} of C2F network with supplementary global matching features $\mathbf{m}^l_{\text{G}}$ by

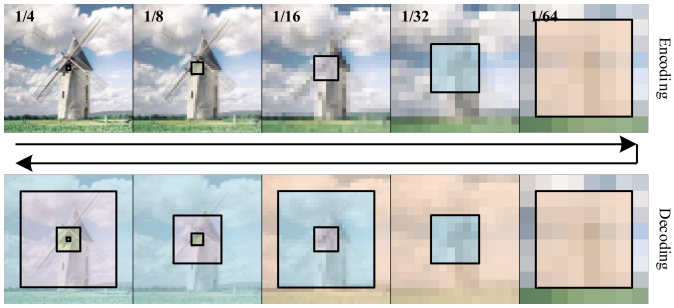

Fig. 4. The Theoretical Receptive Field of GMC. In this sample, the resolution of inputs is 512×512, the colored region is the search range relative to inputs' resolution at each resolution level. The convolution order is in arrow direction.

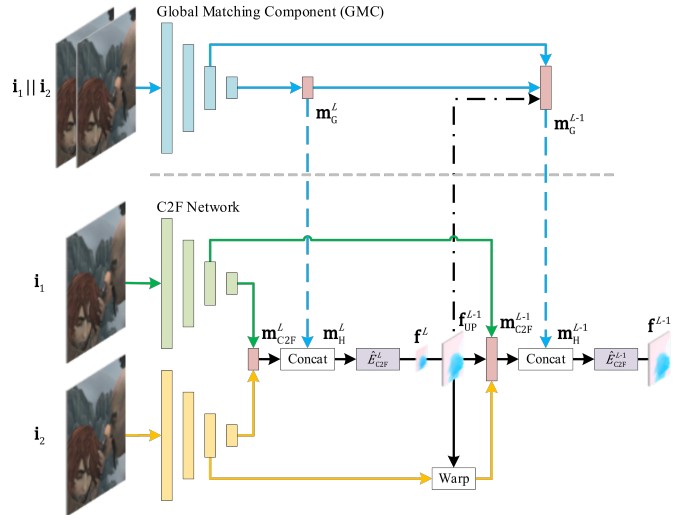

Fig. 5. Hybrid Matching Optical Flow Network (HMFlow). The GMC and C2F network are above and below the grey dotted line respectively.

small calculation cost. And $\mathbf{m}_{\mathrm{G}}^{l}$ will guide E_{C2F}^{l} to discover the missed small and fast-moving objects.

The basic architecture of GMC is described in TABLE I. The $\mathbf{i}_1||\mathbf{i}_2$ is two consecutive images concatenated in channel dimension. And the $\mathbf{m}_{\mathrm{G}}^{l}$ is the output global matching feature of GMC at $1/2^l$ resolution level. Same as encoder-decoder optical flow networks, the search range of GMC is the receptive fields of its convolution layers. As Fig. 4 shown, the theoretical receptive field of GMC expands by encoding, and further expands through skip-connection in decoding. By this way, the search range of $\mathbf{m}_{\mathrm{G}}^{l}$ increases with resolution to whole images, and keep detail information at the same time. This process avoids the dependence of warped features according to up-sampled flows.

C. Hybrid Matching Optical Flow Network (HMFlow)

Taking advantage of large search range of GMC's global matching features and high-quality of C2F network's local matching features, we build new Hybrid Matching Optical Flow Network (HMFlow) to capture the small and fast-moving objects and keep network's high accuracy and small model

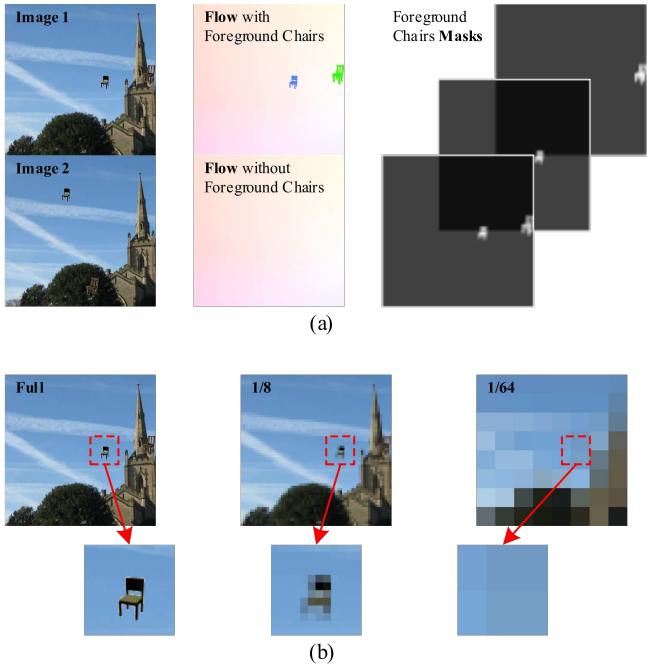

Fig. 6. SFChairs Dataset. (a) is a sample with images, optical flow ground truth, and masks for foreground chairs in SFChairs. (b) indicates the foreground chair at different resolution levels, the chair disappears at the lowest resolution level.

size. This new network in Fig. 5 includes two part, a GMC and a C2F network. HMFlow can be built according to existing C2F network in three steps, (i) network dividing, (ii) component building, and (iii) network integrating.

Network Dividing. Selecting a C2F network and setting its max spatial pyramid level to be L_{C2F}. Dividing this network into feature matching part M_{C2F} and flow estimating part E_{C2F}^{l} according to III-A.

Component Building. Setting GMC's max spatial pyramid level to be $L_{\mathrm{G}} = L_{\mathrm{C2F}}$. Setting GMC's number of channels c_l according to the corresponding one in the Siamese network of selected C2F network.

Component Integrating. The matching features of GMC and selected C2F network are concatenated in channel dimension to hybrid matching features $\mathbf{m}_{\mathrm{H}}^{l}$:

$$\mathbf{m}_{\mathrm{H}}^{l} = \mathbf{m}_{\mathrm{G}}^{l}||\mathbf{m}_{\mathrm{C2F}}^{l} \tag{3}$$

And then, the number of channels of E_{C2F}^{l} are rectified to fit $\mathbf{m}_{\mathrm{H}}^{l}$. The $\mathbf{m}_{\mathrm{H}}^{l}$ are inputed to the rectified $\hat{E}_{\mathrm{C2F}}^{l}$ in place of $\mathbf{m}_{\mathrm{C2F}}^{l}$:

$$\mathbf{f}^{l} = \hat{E}_{\mathrm{C2F}}^{l}(\mathbf{m}_{\mathrm{H}}^{l}, \mathbf{u}_{\mathrm{C2F}}^{l}) \tag{4}$$

Finally, the up-sampled flows $\mathbf{f}_{\mathrm{UP}}^{l}$ are sent to Match$l$ of GMC.

In HMFlow, $\mathbf{m}_{\mathrm{G}}^{l}$ with global search range can guide $\hat{E}_{\mathrm{C2F}}^{l}$ to discover the small and fast-moving objects dismissed by $\mathbf{m}_{\mathrm{C2F}}^{l}$, and $\mathbf{m}_{\mathrm{C2F}}^{l}$ with high-quality can make the network keep high accuracy. Meanwhile, $\mathbf{f}_{\mathrm{UP}}^{l}$ can assist the GMC in more accurate global matching. By this way, HMFlow can realize bi-directional enhancement of GMC and C2F network.

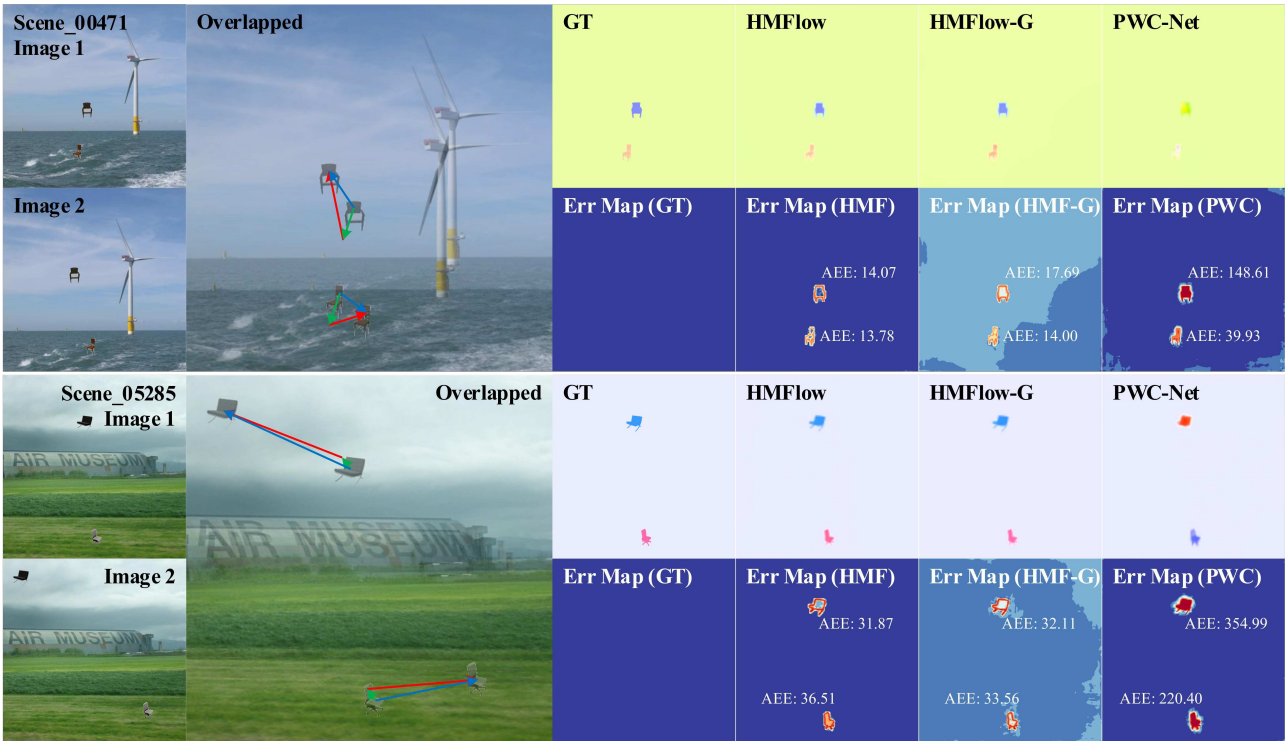

Fig. 7. The Results And Error Maps on The Test Set of SFChairs. The absolute and relative trajectory of foreground objects and the motions of background behind the objects are marked by blue, red, and green arrow line in the overlapped images. The AEEs of foreground chairs are marked in the error maps. The HMFlow-G estimates flows with only GMC's global matching features.

D. Loss

The HMFlow can be trained using the proposed loss functions. Let Θ be the set of all learnable parameters in both GMC and C2F network. Let $\mathbf{w}^l_\Theta$ denote the estimated flow at the lth level, and $\mathbf{w}^l_{\mathrm{GT}}$ the corresponding ground truth. At the pre-training stage, we use the same multi-scale training loss as the one used by [11]:

$$\mathcal{L}(\Theta) = \sum_{l=l_0}^{L} \alpha_l \sum_{\mathbf{x}} |\mathbf{w}^l_\Theta - \mathbf{w}^l_{\mathrm{GT}}|_2 + \gamma|\Theta|_2 \tag{5}$$

where $|\cdot|_2$ computes the L2 norm of a vector and the second term regularizes parameters of the network. For fine-tuning, we use the robust training loss in [15]:

$$\mathcal{L}(\Theta) = \sum_{l=l_0}^{L} \alpha_l \sum_{\mathbf{x}} (|\mathbf{w}^l_\Theta - \mathbf{w}^l_{\mathrm{GT}}| + \epsilon)^q + \gamma|\Theta|_2 \tag{6}$$

where $|\cdot|$ denotes the L1 norm, $q < 1$ gives less penalty to outliers, and ϵ is a small constant.

IV. Small and Fast-Moving Chairs Dataset (SFChairs)

In order to quantitatively and qualitatively evaluate the performance of optical flow networks in this problem, we build a new dataset for small and fast-moving objects, named Small and Fast-Moving Chairs (SFChairs). Fig. 6 (a) shows a sample in SFChairs. The scenes of this dataset are similar to FlyingChairs, but the scales of the foreground objects and the relative motion between foreground and background are elaborate set according to the definition of small and fast-moving objects in [16], which makes all foreground objects become the small and fast-moving objects. Both optical flow ground truth with and without foreground objects is provided. And the pixel-level masks for each foreground object are recorded. We use the chairs from [35] as foreground and select 2798 images from Places [36] as background. As Fig. 6 (b) shown, the scales of the chairs are less than 64 pixels to guarantee that they are small enough to disappear at the top level of spatial pyramid. SFChairs contains 10,000 examples with a resolution of 512×512 that we split into 90% training set and 10% test set.

V. Experiments

Network Details. We use PWC-Net, a representative end-to-end C2F optical flow networks, as a baseline. HMFlow is built according to the scheme in section III-C. (i) Network Dividing, PWC-Net's max spatial level is set to be $L_{\mathrm{PWC}} = 6$, the flow estimating part of PWC-Net E^l_{PWC} is the Optical Flow Estimator at corresponding resolution level, and the feature matching part M_{PWC} is the rest of components. (ii) Component Building, GMC's max spatial pyramid level is set to be $L_{\mathrm{G}} = L_{\mathrm{PWC}} = 6$ and the numbers of channels are set to be $c_1 = 16$, $c_2 = 32$, $c_3 = 64$, $c_4 = 96$, $c_5 = 128$, and $c_6 = 196$ according to PWC-Net's Feature Pyramid

TABLE II
AEEs ON SFCHAIRS

Models	Training Set			Test Set		
	All	Bg.	Obj.	All	Bg.	Obj.
PWC-Net	(0.62)	(0.27)	(64.54)	0.79	0.27	87.01
HMFlow-G	(0.59)	(0.36)	(45.58)	0.71	0.42	56.03
HMFlow	(**0.39**)	(**0.20**)	(**36.64**)	**0.45**	**0.21**	**44.34**

[a] The **All**, **Bg.** and **Obj.** indicate the AEEs of All image, Background and Foreground Object Regions.
[b] The **HMFlow-G** estimates flows with only GMC's global matching features.

Extractor. (iii) Component Integrating, the hybrid matching features $\mathbf{m}_{\mathrm{H}}^{l}$ is made by 3D Cost Volume features of PWC-Net and $\mathbf{m}_{\mathrm{G}}^{l}$ of GMC, the up-sampled flows of PWC-Net are sent to corresponding Match$l$ of GMC.

Training Details. We build and train networks in PyTorch [37]. The weights in the training losses, Eq. 5 and 6, are set to be $\alpha_6 = 0.32$, $\alpha_5 = 0.08$, $\alpha_4 = 0.02$, $\alpha_3 = 0.01$, and $\alpha_2 = 0.005$. The weight γ of L2 penalty is set to be 0.0004. First, we pre-train HMFlow on FlyingChairs and FlyingThings3D according to learning rate schedule S_{long} and S_{fine} [12]. Then, we fine-tune network on SFCharis and MPI Sintel.

A. SFChairs

We fine-tune the pretrained HMFlow and PWC-Net on SFChairs with same learning rate schedule.

Quantitative Analysis. Quantitative evaluation of networks performance based on Average End-Point Error (AEE), which represents the Euclidean distance between the estimated results and the ground truth. In TABLE II, we compare the AEEs of HMFlow with PWC-Net on SFChairs. The performance of HMFlow has distinct advantage over PWC-Net on both training set and test set. The AEEs of slow-moving background are similar for both networks. However, drawing the global matching features from GMC, HMFlow's AEEs of small and fast-moving foreground objects are about 50% of PWC-Net. It proves that the improvement effect of HMFlow on this problem in original PWC-Net is obvious.

Qualitative Analysis. Fig. 7 compares the results of HMFlow with PWC-Net on the test set of SFChairs. HMFlow can capture the small and fast-moving chairs in the scenes, while their motion is mis-estimated by PWC-Net.

Fig. 8 exhibits how the global matching features of GMC play a role in the spatial pyramid of the C2F network. The chair in the center of the input images can't be seen clearly in the images with $1/64$ and $1/32$ resolution levels. And at $1/16$ resolution level, when the second image is warped according to the up-sampled optical flow, this chair moves with the background and goes beyond the local search range of $\mathbf{m}_{\mathrm{C2F}}^{4}$. PWC-Net's flow estimating part E_{PWC}^{4} just depends on this local matching feature $\mathbf{m}_{\mathrm{PWC}}^{4}$ whose search range is unable to cover this chair. Hence, this chair's motion will not be able to correctly estimated at current and following resolution levels. In contrast, all $\hat{E}_{\mathrm{C2F}}^{l}$ of HMFlow can get the supplementary

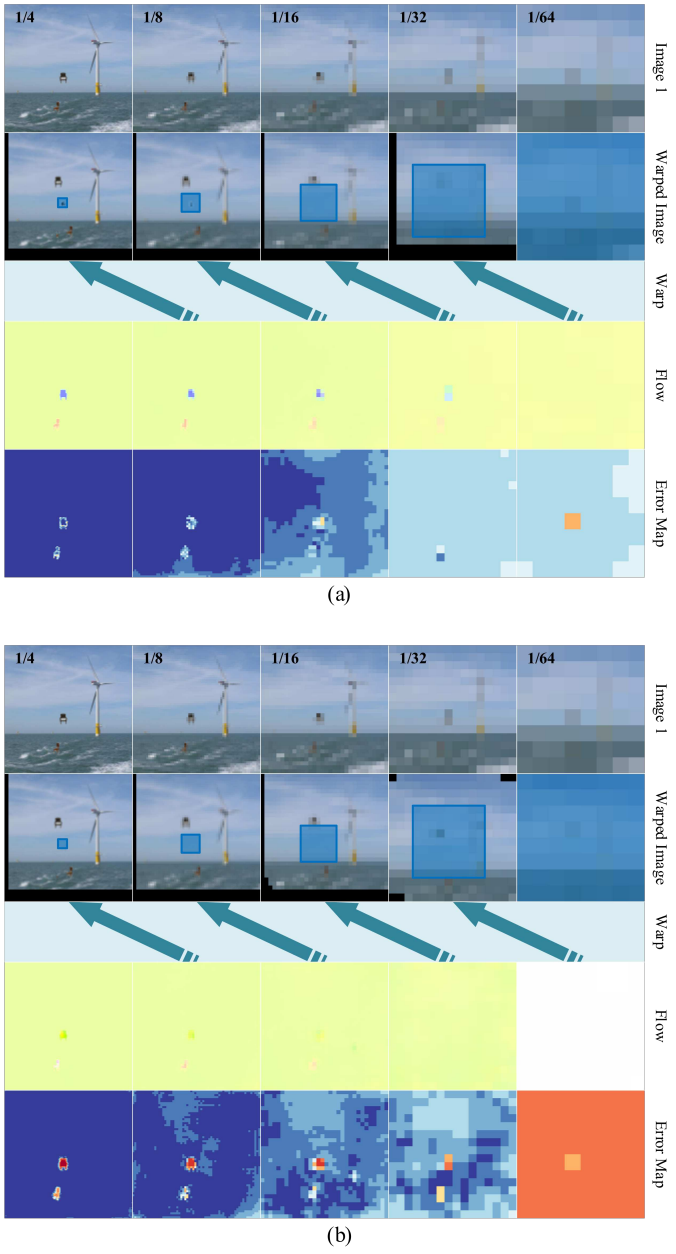

Fig. 8. The Results in Spatial Pyramid. (a) and (b) show the results of HMFlow and PWC-Net. The blue regions in the warped images indicate the local search ranges of $\mathbf{m}_{\mathrm{C2F}}^{l}$.

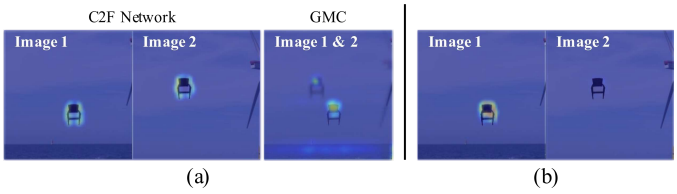

Fig. 9. The Saliency Maps. The saliency maps visualise the most concentrated regions of the foreground chair in inputs. (a) and (b) are the maps of HMFlow and PWC-Net.

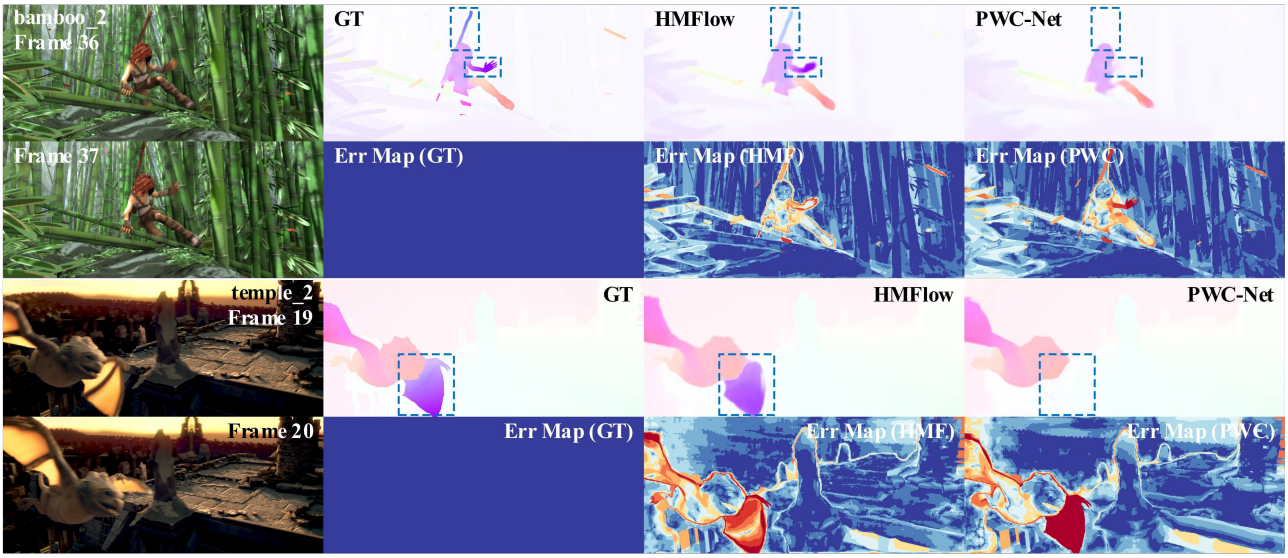

Fig. 10. The Results and Error Maps on The Training Set of MPI Sintel. The small and fast-moving objects are marked with blue boxes in optical flows.

global matching information $\mathbf{m}_{\mathrm{G}}^{l}$ from GMC. By $\mathbf{m}_{\mathrm{G}}^{4}$, this chair's flow can be estimated by E_{C2F}^{4} at $1/16$ resolution level, and is warped back to the local search range of subsequent of $\mathbf{m}_{\mathrm{C2F}}^{3}$ and of $\mathbf{m}_{\mathrm{C2F}}^{2}$.

We visualise saliency maps [38] to show the divertion of the C2F network's attention before and after the GMC introduction. Fig. 9 demonstrates the attentive location of the small and fast-moving chair in the inputs. Limited by local search range, PWC-Net can just pay attention to the chair in the first image in (b). It means that PWC-Net fails to locate this chair from the second image. In contrast, the GMC of HMFlow is able to notice the chair before and after moving at the same time, and guide the C2F network pay attention to the chair in the second image, and match it with the one in the first image.

Ablation Study. To analyze GMC's global matching features, we run an ablation on HMFlow. HMFlow-G removes M_{C2F}'s all local matching features $\mathbf{m}_{\mathrm{C2F}}^{l}$, but keeps the unmatching features $\mathbf{u}_{\mathrm{C2F}}^{l}$. All $\hat{E}_{\mathrm{C2F}}^{l}$ of HMFlow-G only use GMC's global matching features $\mathbf{m}_{\mathrm{G}}^{l}$ to estimate optical flows.

In TABLE II, HMFlow-G's AEEs of foreground objects on training set and test set are about 65% and 70% of PWC-Net, while its AEEs of background are higher. The results in Fig. 7 also reflect same the same phenomenon. This ablation shows that GMC's global matching features can effectively match the small and fast-moving objects and help network to estimate their motion. But these lightweight global matching features perform obviously poorly in the large smooth regions, *e.g.*background. And our full HMFlow can complement the local and global matching features, and achieves high performance in all regions.

B. MPI Sintel

We fine-tune the pre-trained HMFlow on MPI Sintel dataset. The robust loss function in Eq. 6 is used with $\epsilon = 0.01$ and $q = 0.4$ [15]. We use the clean and final passes of this training data throughout the fine-tuning process.

TABLE III
AEEs ON MPI SINTEL

Methods	Training Set		Test Set		Size
	Clean	Final	Clean	Final	(million)
FlowNetS [11]	(3.66)	(4.44)	6.96	7.76	38.68
FlowNetC [11]	(3.78)	(5.28)	6.85	8.51	39.18
FlowNet2 [12]	(1.45)	(2.01)	4.16	5.74	162.52
SPyNet [13]	(3.17)	(4.32)	6.64	8.36	1.20
LiteFlowNet [14]	(**1.35**)	(**1.78**)	4.54	5.38	5.37
PWC-Net [15]	(1.70)	(2.21)	3.86	5.13	9.37
HMFlow	(1.44)	(2.23)	**3.21**	**5.04**	14.27

[a] The **Size** indicates networks' number of parameters in million.

Accuracy and Model Size. TABLE III compares HMFlow with the representative encoder-decoder [11], [12] and C2F [14], [15] optical flow networks. The AEEs of HMFlow are lower than all networks participating in the comparison on MPI Sintel test set. Due to the use of GMC, HMFlow's model size is slightly larger than its baseline, PWC-Net. Even though, its parameters are just $1/3$ of FlowNet [11] and $1/11$ of FlowNet 2.0 [12]. The results show that our HMFlow can take the advantage of C2F network's small model size and relative high accuracy.

Small and Fast-Moving Objects. We qualitatively compare the results of HMFlow with PWC-Net on MPI Sintel training set, which provides ground truth. There are some small and fast-moving objects in the scenes in Fig. 10, such as the swinging arm, slender stick, and fanning wing. HMFlow can estimate flows of these objects accurately, while PWC-Net dismisses their motion. This shows that our HMFlow is effective for this problem in general situations, instead of

overfitting on SFChairs.

VI. Conclusion

In this paper, we have studied the problem of capturing the small and fast-moving objects in C2F optical flow networks. To address the mismatching of C2F networks' warping based local matching features, we have designed a lightweight Global Matching Component (GMC) for global matching features. We have built new Hybrid Matching Optical Flow Network (HMFlow) by integrating GMC into existing C2F networks. We have built a specialized dataset for small and fast-moving objects, SFChairs. The experiments have proved the HMFlow's effectiveness that our network can solve the problem of capturing small and fast-moving objects with small model size and high accuracy.

References

[1] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36, no. 6, pp. 1187–1200, 2014.

[2] Y.-H. Tsai, M.-H. Yang, and M. J. Black, "Video segmentation via object flow," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[4] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[5] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[6] G. Costante and T. A. Ciarfuglia, "Ls-vo: Learning dense optical subspace for robust visual odometry estimation," *The IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1735–1742, 2018.

[7] F. C. Glazer, "Hierarchical motion detection," 1987.

[8] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *International Journal of Computer Vision*, vol. 2, no. 3, pp. 283–310, 1989.

[9] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *The European Conference on Computer Vision (ECCV)*. Springer, 1992, pp. 237–252.

[10] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *The European Conference on Computer Vision (ECCV)*. Springer, 2004, pp. 25–36.

[11] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[12] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2017, p. 6.

[13] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] T.-W. Hui, X. Tang, and C. C. Loy, "Liteflownet: A lightweight convolutional neural network for optical flow estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[15] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[16] T. Brox and J. Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *The IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 33, no. 3, pp. 500–513, 2011.

[17] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[18] M. J. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer vision and image understanding*, vol. 63, no. 1, pp. 75–104, 1996.

[19] D. Shulman and J.-Y. Herve, "Regularization of discontinuous flow fields," in *Visual Motion, 1989., Proceedings. Workshop on*. IEEE, 1989, pp. 81–86.

[20] A. Bruhn and J. Weickert, "Towards ultimate motion estimation: Combining highest accuracy with real-time performance," in *The IEEE International Conference on Computer Vision (ICCV)*, vol. 1. IEEE, 2005, pp. 749–755.

[21] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 9, pp. 1744–1757, 2012.

[22] T. Brox, C. Bregler, and J. Malik, "Large displacement optical flow," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 41–48.

[23] Y. Hu, R. Song, and Y. Li, "Efficient coarse-to-fine patchmatch for large displacement optical flow," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] T.-W. Hui, X. Tang, and C. C. Loy, "A Lightweight Optical Flow CNN - Revisiting Data Fidelity and Regularization," 2020. [Online]. Available: http://mmlab.ie.cuhk.edu.hk/projects/LiteFlowNet/

[25] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Models matter, so does training: An empirical study of cnns for optical flow estimation," *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, to appear.

[26] Y. Lu, J. Valmadre, H. Wang, J. Kannala, M. Harandi, and P. Torr, "Devon: Deformable volume network for learning optical flow," in *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 2705–2713.

[27] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[28] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *The European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 611–625.

[29] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3354–3361.

[30] R. Li, R. T. Tan, and L.-F. Cheong, "Robust optical flow estimation in rainy scenes," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[31] J. Janai, F. Güney, A. Ranjan, M. Black, and A. Geiger, "Unsupervised learning of multi-frame optical flow with occlusions," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[32] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep joint image filtering," in *The European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 154–169.

[33] ——, "Joint image filtering with deep convolutional networks," *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 8, pp. 1909–1923, 2019.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[35] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, "Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[36] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 6, pp. 1452–1464, 2017.

[37] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[38] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *Computer Science*, 2013.